Seabed classification using a bag-of-prototypes feature representation

Timm Schoening[1], Thomas Kuhn[2], and Tim W Nattkemper[1]

[1] Biodata Mining Group, Bielefeld University, Germany
tschoeni@cebitec.uni-bielefeld.de
[2] Institute for Geosciences and Natural Resources (BGR), Hanover, Germany

Abstract. The increasing scientific and economic interest in the visual exploration and monitoring of marine areas is creating huge amounts of new underwater image and video data and approaches to computationally assisted analysis are desperately needed. In this paper we propose an image patch feature representation concept, the Bag of Prototypes (BoP), to cope with the individual problems in underwater image analysis. We consider the case of seafloor classification, which is relevant in many contexts such as habitat mapping or the exploration of mineral resources and show, that the BoP concept allows an efficient and accurate tile-wise estimation of poly-metallic nodule coverage in relation to two differently acquired gold standards.

Keywords: Underwater image analysis, Feature Representation, Resource Exploration

1 Introduction

More than seventy percent of the earth's surface are covered with water. The sheer size of this domain calls for automated exploration technologies and this features some unique challenges for computer vision (CV) research. Shallow coastal waters are difficult to explore visually as biological and geological factors can create massive turbidity. Still more than sixty percent of earth's surface lie in the aphotic zone below 200m depths, i.e. where no sunlight reaches. As a consequence, biological factors become less influential to image quality and the geophysical properties remain relatively stable throughout the year.
Technologies to visually explore the marine areas range from remotely operated vehicles (ROV) via ocean floor observation systems (OFOS) to automated underwater vehicles (AUV). All of these create rapidly growing piles of image data, leading to serious bottlenecks in analysis [6]. The image analysis aims at i) classifying an image (e.g. in habitat mapping [11]), ii) image segmentation (e.g. in exploration [13]) or iii) detect singular object instances (e.g. in biodiversity studies [10]).
In the deep sea domain, the illumination stems from the artificial light source of the camera. Still the illumination and captured area is not completely controllable, due to the relative movement of the camera rig. This factor requires the normalisation of the captured image data in a pre-processing step.
The second challenge concerns the tuning of any pattern recognition algorithm where

it is applied to i) - iii). It has been reported that it is difficult to gather manual annotations i.e. by manually exploring and tagging the data (a gold standard for training and validation) [6]. This is partly due to the scarce occurrence of objects of interest which often results in erroneous annotations. Also, the trend to big data has reached the marine scientific field (e.g. in terms of growing image resolution) and manual data analysis is usually carried out by single researchers rather than well-organised groups [7], which makes the creation of a detailed, fully annotated training set infeasible.
In this paper, we will consider a typical exploration scenario: the quantification of poly-metallic nodule coverage in underwater image transects recorded with an OFOS. The images show the seabed, recorded with an HD camera in a top-down view. The seabed is covered with objects of a given class (here poly-metallic nodules) and the objects show variations in their features due to their size and coverage with sediments [15]. Similar tasks in underwater image analysis are small-scale habitat mapping [16] or bacterial mats classification [5]. Comparable binary segmentation tasks from other domains can be similarly approached. In this paper, we use a Bag-of-Words (BoW) based approach [14] which is applied in a tile-wise regression concept to estimate the degree of nodule coverage. The BoW method is referred to as Bag-of-Prototypes here, as low-level feature representations of image pixels are mapped to cluster prototypes. This mapping is done with an H^{2}SOM to allow for a more complex tessellation of the feature space than the classical k-Means method in BoW could provide.
The BoP (and similarly the BoW) approach is motivated as follows: In pattern recognition, the basic task is to find a function f that maps an entities' feature representation $\mathbf{x}_i$ to an output: $f(\mathbf{x}_i) \mapsto w_*$ where w_* can be a distinct category (classification) or a quantitative output (regression). The feature vector $\mathbf{x}_i$ describes (the neighbourhood of) a pixel. The function f can for instance be approximated with machine learning methods like Random Forests (RF) or Support Vector Machines (SVM).
Due to the annotation problems, supervised learning algorithms that learn f directly from the training data could not be applied which leaves unsupervised approaches like learning vector quantisation [3]. Here we will consider the Hierarchical Hyperbolic Self-Organizing Maps (H^{2}SOM) [9]. The final mapping is achieved using the prototype vectors $\mathbf{u}_j$ estimating the data distribution of all $\mathbf{x}_i$ in the feature space:

$$\mathbf{x}_i \mapsto \mathbf{u}_j \mapsto w_* \tag{1}$$

In well-separable cases, both these mappings can be done unambiguously. But in real-world data, such as underwater images this is not the case. To adapt the approach, the two mappings (Equation 1) are interpreted less deterministic, either the first P($\mathbf{u}_j|\mathbf{x}_i$) $\in [0..1]$which is often referred to as the fuzzy method, or a non-deterministic association of each prototype to a class: P($\mathbf{u}_j, w_*$) $\in [0..1]$.
However, we observe that in underwater images, the objects of interest do not show the tendency to distribute their features in a small set of cluster prototypes but do display specific heterogeneous combinations of a large number of matching prototypes which is often the case in underwater imaging due to coverage with sediments, coral rubble etc. This led to the development of the presented bag-of-prototypes (BoP) approach.

2 Materials and Methods

The images were taken in ∼6,000 m depth in the Pacific Ocean and show a top-down view of the seafloor which is covered with resource rich poly-metallic nodules (see Figure 1). The interest is to estimate the total nodule coverage on the seafloor. The nodules are of different size and show a varying morphology and thus pattern matching approaches will not suffice. The images were captured with an OFOS that was steered to hover about two to four meters above the seafloor. All images feature an illumination cone caused by a single flash light source that could not illuminate the seafloor continuously. A fundamental challenge in underwater imaging can be seen towards the corners of the images (see Figure 1): the signal is darker and blurry here. This is caused by the physical properties of water where longer lights paths lead to increased scattering and absorption. Also, a wide-angle lens was used that further increased this effect. For computational speedup, a 2,000 × 1,000 pixel region was selected in each of the N_I = 334 images. Only within these regions, feature vector were extracted and coverage estimates computed. On average, these centre regions contain 308 poly-metallic nodules (which has been estimated using manual evaluations, see below).
Despite the general in expert annotation, two methods of creating a gold standard were applied here. From the traditional manual exploration, binary masks existed that were based on a hand-tuned thresholding of the intensities within each individual image I. This method utilised the originally captured images without pre-processing and was tuned in a way that overestimates the visible nodule coverage. This approach is very time-consuming since the threshold was tuned for each image and is error-prone as it assumes a constant illumination throughout the image. Still it provides the most detailed manual annotation possible. Due to the drawbacks of creating such a pixel-perfect annotation, a further method was implemented where the images were cut into 4 × 8 virtual annotation tiles T_g. An expert in visual nodule exploration annotated each T_g, within a set of $N_S = 9$ random sample images with a coverage estimate c_g^T in steps of ten percent: i.e. $c_g^T \in \{0, 10, 20, ..., 100\} (g = 0..N_T - 1)$ with $N_T = N_S \times 8 \times 4 = 288$. This annotation was done with the web-based image annotation software Biigle [8]. To make the the binary masks comparable to the c_g^T, the masks were similarly cut to tiles of the same size and the amount of nodule-positive pixels counted within each tile. This led to coverage estimates c_g^M.

Pre-processing: Two steps are required to normalise the image signal. First, to spatially normalise the illumination within an image the illumination cone induced by the artificial light source is removed by creating a subtraction image I^S. Second, to normalise the images within the transect to a similar colour spectrum, the intensity histogram of I^S is gamma-corrected such that the peak lies in the middle of the spectrum. A detailed explanation of the pre-processing is available in [12] together with an approach to tune the pre-processing parameters automatically.
For comparison, we also applied, e.g. the Gray World algorithm [2] that showed inferior performance, probably due to the unusual colour characteristics of underwater images that do not fulfil the assumptions made. Another method developed especially for underwater images [1] tended to produce massive colour shifts that complicated a visual interpretation. The resulting images I (see Figure 1) feature an enhanced colour spec-

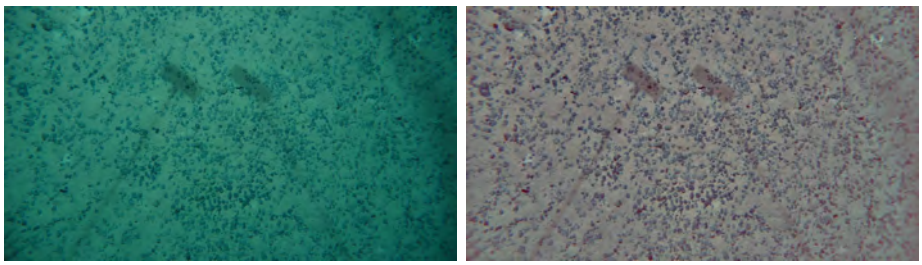

Fig. 1. A seafloor image and the result after the colour pre-processing.

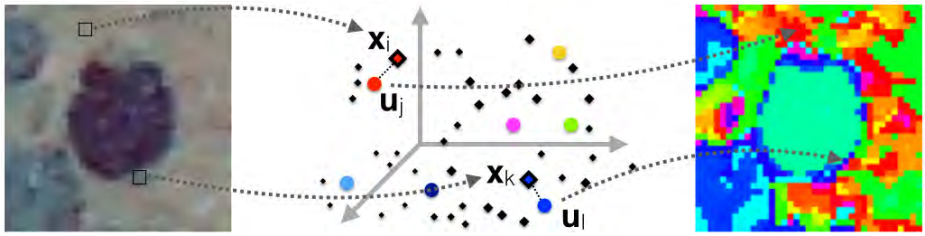

Fig. 2. The H^{2}SOM index image: First, RGB histogram features vectors ($\mathbf{x}_i, \mathbf{x}_k \in \mathbb{R}^{48}$) are extracted from I. From these features, the BMU among all $\mathbf{u}_j$ of the H^{2}SOM (coloured circles) is determined (middle graphic in a three dimensional representation). The index of the pixel's BMU is then stored in a new image I^u at the same location. Due to the H^{2}SOM topology, these indices can be colour coded (right part) with a topology preserving colour code.

trum, homogenised contract between nodules and sediment and reduced illumination cone.

Feature mapping: From the pre-processed image I, colour histogram features are extracted. A 48-dimensional feature vector $\mathbf{x}_i$ is constructed for each pixel p_i of the images. This feature vector consists of the frequencies of the colours in a $\theta_1 \times \theta_1$ pixel neighbourhood around the feature vector's corresponding pixel. The 256 intensity values of each of the RGB colour channels were mapped to 16 equally sized bins ($16 \times 3 = 48$D). The application of colour features was only possible due to the image normalisation. The inclusion of texture features (e.g. edge histograms or Gabor wavelets) yielded inferior regression performance.

To cluster the feature vectors, an H^{2}SOM was used that consisted of $\theta_2 = 161$ prototypes ($j = 0..\theta_2 - 1$). We chose to use the H^{2}SOM as it has a good learning-performance, as well as a fast way of finding the best-matching unit (BMU) for a new data sample (beam search). A BMU image I^u was created for each I. Therefore each pixel of I was set to the prototype index of the pixel's feature vector's best matching prototype of the H^{2}SOM (e.g. j if $\mathbf{u}_j$ was the BMU, i.e. BMU($\mathbf{x}_i$) = $\mathbf{u}_j$). That way an index image (or cluster map) I^u is created (see Figure 2). Due to the topology preserving, low dimensional embedding of the H^{2}SOM prototypes, these indices can be attributed with a colour value, based on the prototypes location on the hue disc of the

HSV colour space [4] (see Figure 2 right).
Manual inspections of these visualisations showed, that about twenty percent of the prototypes could mostly be assigned to either the sediment or nodule category, while the others could not reliably be assigned to one of those categories. First, there were several "transitional" prototypes at the nodule margins. Second, and more challenging, some prototypes occurred at various singular locations within the background, object and transitional regions.

Bag-of-prototypes representation: The central idea of the BoP approach is to integrate the pixel neighbourhood to the mapping in Equation 1:

$$\{\mathbf{x}_i\} \mapsto \{\mathbf{u}_j\} \mapsto \mathbf{b}_i \mapsto w_* \tag{2}$$

The mapping from $\mathbf{x}_i$ to $\mathbf{u}_j$ is done deterministically, in this case with the H^{2}SOM. A set of prototypes $\{\mathbf{u}_j\}$ is then grouped to the additional feature representation $\mathbf{b}_i$ as follows: The basis is I^u in which the prototype frequencies b_i^l within a squared, $\theta_3 \times \theta_3$ pixel neighbourhood around a pixel p_i are computed. This means that the $\mathbf{b}_i$ belonging to pixel p_i contains a frequency count of all prototype indices l occurring in the θ_3^2 neighbourhood of p_i in I^u. That way, local distributions of prototypes are characterised that only together represent the setup of a visual pattern (e.g. the nodules).
The common BoW approach works similarly but is usually applied to whole images for retrieval tasks. Here it is used in a regression task for small-scale image representation to estimate resource coverage.

Coverage estimate: The final mapping ($\mathbf{b}_i \mapsto w_*$) still requires semantic input e.g. in form of the expert annotations. To test the BoP representation $\mathbf{b}_i$, the k-Nearest-Neighbor algorithm was used as a straightforward reference classifier. The c_g^T and c_g^M served as the gold standard in a leave-one-out strategy and yielded coverage estimates $c_g^{\text{BoP}/M}$ and $c_g^{\text{BoP}/T}$. Of course the application of more advanced learners (e.g. Random Forests, Support Vector Machines) can be considered to further improve the regression performance. Nevertheless, in this work we wanted to focus on the BoP approach.
To compare the results obtained via the $\mathbf{b}_i$ with the classical way of assigning the $\mathbf{u}_j$ to either w_{nodule} or w_{sediment} we tried to find the optimal separation of these classes manually. Anyway this assignment is subjective and we thus created three different selections, with varying degrees of confidence in the prototypes (i.e. a conservative set Γ_{c} ($|\Gamma_{\text{c}}| = 12$), a medium set Γ_{m} ($|\Gamma_{\text{m}}| = 21$) and a liberal set Γ_{l} ($|\Gamma_{\text{l}}| = 26$). From these selections, coverages c_g^c, c_g^m and c_g^l were computed similar to c_g^M.
The parameters $\theta_i, i = 1, 2, 3$ influence the BoP approach and can be tuned for a specific CV challenge.

3 Results

Figure 3 shows four resulting $\mathbf{b}_i$ for different patches of the sample image. Three nodules are highlighted as well as a sediment section. The $\mathbf{b}_i$ corresponding to the pixel neighbourhoods are shown and differ in all cases (see caption for more details).

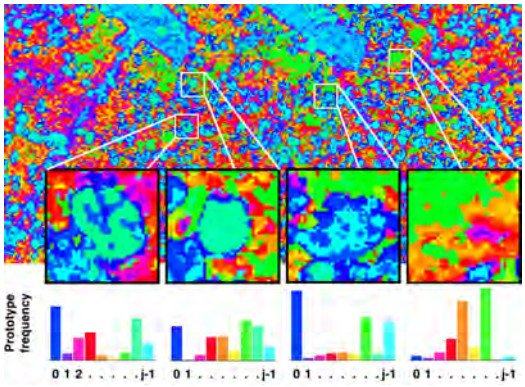

Fig. 3. The effect of the prototype mapping is shown. Four pixel neighbourhoods or patches have been highlighted that stand for the square pixel neighbourhoods. For visualisation purposes we chose $\theta_3 = 36$. The first three patches show poly-metallic nodules, the last a pure sediment patch. Below, the corresponding $\mathbf{b}_i$ are shown as histograms. Shown are only nine prototypes and all occurring b_i^l are mapped to the closest of these, regarding the HSV colour. Nodules appear "blueish/turquiose", but an unambiguous assignment of $\mathbf{u}_j$ to w_* is not possible as all $\mathbf{u}_j$ occur in all four patches. Also, around the nodules exist further "greenish/blueish" regions, that lie within the sediment region of the image.

The total error was measured by taking the absolute of the tile-wise difference between two coverage estimates c_g^α and c_g^β. This value was normalised to produce an average per-image error:

$$E(\alpha, \beta) = \frac{1}{N_T} \sum_{n=0}^{N_T - 1} |c_g^\alpha - c_g^\beta| \tag{3}$$

The errors compare the coverage estimates with the gold standard and thus give the deviation in percentage points per image. Figure 4 (left) shows $E(\alpha, \beta)$ for six different experiments (see caption for details). By hand tuning the prototype selection, we were able to obtain a set (Γ_l) that outperformed the BoP approach ($E(T, l) = 3.72$ vs $E(T, \text{BoP}/T) = 4.12$ percent points). Anyway the creation of each Γ was time-consuming and subjectively while the BoP approach does not require any prototype specific assumptions. The medium set produced a slightly higher error than the BoP approach ($E(T, m) = 4.93$) while the conservative set Γ_c produced the highest error ($E(T, c) = 9.68$).

The three parameters θ_1, θ_2 and θ_3 control the BoP approach. The first two were set to $\theta_1 = 7$ and $\theta_2 = 161$ heuristically. The effect of other settings will be evaluated in the future. We looked at the effect of different θ_3 and found lower errors $E(\alpha, \beta)$ for larger θ_3 (see Figure 4, right). This reflects the effect, that for larger tiles, the BoP features become less variable for single coverage categories. As the resource mining, if ever, will take place on a scale larger than the images, this is beneficial, especially as larger θ_3 lead to shorter computation time. During some trials, we allowed negative deviations from the gold standard values rather than using absolute values only (which we did as we wanted to present the worst case performance here). This led to lower

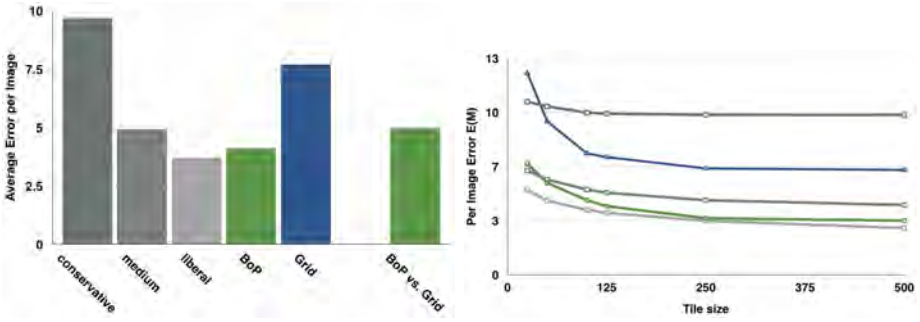

Fig. 4. To the left, the average per-image errors for six settings.The first five columns are all obtained with the c_g^T as a gold standard. The first three columns show the manually selected sets $\Gamma_c, \Gamma_m, \Gamma_l$. The fourth column shows the BoP result $E(T, \text{BoP}/T)$ and the fifth the mask annotation c_g^M compared to the tile annotation c_g^T (i.e. $E(T, M)$). Here it can be seen, that the grid annotations differ from the fine-scale annotations c_g^T. The sixth column shows the BoP result with the c_g^M as the gold standard (i.e. $E(M, \text{BoP}/M)$). From the columns four and six it can be seen, that the BoP can describe the tile's feature setup qualitatively and is able to match similar tiles (i.e. tiles with similar nodule coverage). Although c_g^T differs from c_g^M, the errors for both BoP trials are low, but it is not clear, which annotation gold standard is better. To the right, the average image-wise error $E(\alpha, \beta)$ vs. tile size θ_3 are shown. Increasing θ_3 leads to smaller per-image errors. The colours of the curves are the same as in the bar chart to the left. Only the $E(T, \beta)$ are given.

errors, as errors of tiles within the same image can average out.
If we look at the mismatch between the gold standards c_g^M and c_g^T we see the ability of BoP to describe the nodule coverage of squared tiles (see Figure 4, left). The problem lies within the semantic annotation which any classification or regression relies upon. After seeing the mismatch between c_g^T and c_g^M we compared the individual tiles and found that while the c_g^M overestimated the nodule coverage, the c_g^T estimates were even larger. This again shows the difficulties in manual annotation without previous training (in case of the c_g^T) but also the subjectivity of any annotation process. As both annotations tend to overestimate the nodule coverage, the coverages determined by BoP will also overestimate the true amount.
In the future, we will apply the BoP approach to other underwater CV problems as well as assess its applicability to other domains.

4 Conclusion

We present the application of the Bag-of-prototypes feature representation to the case of poly-metallic nodule exploration. The approach is capable of describing the local feature setup of complex object entities (i.e. the nodules) while it is straightforward to implement. The advantage over the standard BoW approach is due to the computational simplicity. First, low-level colour features are used, rather than the common SIFT/SURF features. Second, the H^{2}SOM is applied here, that allows for a more efficient BMU search compared to the classical k-Means. Finally, obtaining gold standard

annotations for the regression step is simplified as tiles are annotated rather than pixels. The BoP approach has the potential to be applied in other (underwater) CV scenarios and is currently investigated in the context of multi-class object detection.

Acknowledgements: We thank the Institute for Geosciences and Natural Resources (BGR) for providing us with the sample images. This research was funded by the German Federal Ministry for Economics and Technology (BMWI, FKZ 03SX344A).

References

1. Stephane Bazeille, Isabelle Quidu, Luc Jaulin, Jean-Philippe Malkasse, et al. Automatic underwater image pre-processing. *Proceedings of CMM'06*, 2006.
2. Arjan Gijsenij and Theo Gevers. Color constancy using natural image statistics and scene semantics. *Pattern Analysis and Machine Intelligence*, 33(4):687–698, 2011.
3. Teuvo Kohonen. *Learning vector quantization*. 2001.
4. Jan Kölling, Daniel Langenkämper, Sylvie Abouna, Michael Khan, and Tim W Nattkemper. Whide–a web tool for visual data mining colocation patterns in multivariate bioimages. *Bioinformatics*, 28(8):1143–1150, 2012.
5. Andree Lüdtke, Kerstin Jerosch, Otthein Herzog, and Michael Schlüter. Development of a machine learning technique for automatic analysis of seafloor image data: Case example, pogonophora coverage at mud volcanoes. *Computers & Geosciences*, 39:120–128, 2012.
6. Norman MacLeod and Phil Culverhouse. Time to automate identification. *Nature*, 467(7312):154–5, 2010.
7. Tim Wilhelm Nattkemper. Are we ready for science 2.0? 2012.
8. J. Ontrup, N. Ehnert, M. Bergmann, and T.W. Nattkemper. BIIGLE - Web 2.0 enabled labelling and exploring of images from the Arctic deep-sea observatory HAUSGARTEN. In *OCEANS 2009*, pages 1–7, 2009.
9. Jörg Ontrup and Helge Ritter. Hyperbolic self-organizing maps for semantic navigation. *Advances in neural information processing systems*, 14(14):1417–1424, 2001.
10. Autun Purser, Melanie Bergmann, Tomas Lundälv, Jörg Ontrup, and Tim W Nattkemper. Use of machine-learning algorithms for the automated detection of cold-water coral habitats: a pilot study. *Marine Ecology Progress Series*, 397, 2009.
11. Paul Rigby, Oscar Pizarro, and Stefan B Williams. Toward adaptive benthic habitat mapping using gaussian process classification. *J of Field Robotics*, 27(6):741–58, 2010.
12. Timm Schoening, Melanie Bergmann, Jörg Ontrup, James Taylor, Jennifer Dannheim, Julian Gutt, Autun Purser, and Tim W Nattkemper. Semi-automated image analysis for the assessment of megafaunal densities at the arctic deep-sea observatory hausgarten. *PloS one*, 7(6):e38179, 2012.
13. Núria Teixidó, Anton Albajes-Eizagirre, Didier Bolbo, Emilie Le Hir, Montserrat Demestre, Joaquim Garrabou, Laurent Guigues, Josep Maria Gili, Jaume Piera, Thomas Prelot, et al. Hierarchical segmentation-based software for cover classification analyses of seabed images (seascape). 2011.
14. Chih-Fong Tsai. Bag-of-words representation in image annotation: A review. *ISRN Artificial Intelligence*, 2012, 2012.
15. Ulrich von Stackelberg and Helmut Beiersdorf. The formation of manganese nodules between the clarion and clipperton fracture zones southeast of hawaii. *Marine Geology*, 98(2):411–423, 1991.
16. Stefan B Williams, Oscar Pizarro, Michael Jakuba, and Neville Barrett. Auv benthic habitat mapping in south eastern tasmania. In *Field and Service Robotics*, pages 275–84, 2010.